What makes LLM-generated Answers Valuable for Research Problem Reframing?

Abhinav Sood
University of Sydney
Sydney, Australia
abhinav.sood@sydney.edu.au
(corresponding author)

Stephen Wan
Data61, CSIRO
Sydney, Australia
stephen.wan@data61.csiro.au

Cecile Paris
Data61, CSIRO
Sydney, Australia
cecile.paris@data61.csiro.au

Kazjon Grace
University of Sydney
Sydney, Australia
kazjon.grace@sydney.edu.au

Abstract

Problem reframing can be understood as a form of p-transformational creativity; as a process, it uncovers frames that make an individual rethink their conceptual space, making new solutions possible. While problem reframing is central to design and other forms of creative practice, such reconceptualisation would also be valuable in research disciplines when conflicting evidence or unexpected findings arise. However, LLMs have already been shown to fail at supporting problem reframing in design contexts. This failure stems from two issues: problem frames require deep contextual understanding, and LLM-generated frames lack the novelty needed to serve as foundations. This paper takes a first step toward understanding how LLMs can contribute to problem reframing in research through a two-stage user study with 7 research students across diverse research disciplines. It concludes by identifying two distinct answer profiles and a set of criteria for evaluating whether AI-generated answers have the potential to support problem reframing in research.

Introduction

Research is generally understood as a method to generate, refine, and extend knowledge about a phenomenon of interest. Accounts of the research process describe it as structured and systematic (Creswell and Creswell 2017; Leedy and Ormrod 2001). For example, Leedy and Ormrod (2001) propose that research follows a cyclical seven-step process. In practice, the execution of these steps is not as linear; researchers tend to seek information from their surrounding context in a non-linear fashion (Foster 2004).

Within such a complicated process, multiple questions arise. In design, the evolution of such questions and design goals through problem reframing is deliberately encouraged. Problem reframing involves redefining a problem by altering the assumptions or perspectives that underlie it to arrive at solutions to the problem that might have otherwise been unreachable (Dorst 2015b). Problem reframing is considered core to the practice of design research, so much so that, as a skill, it has allowed designers to create value across disciplines and support transdisciplinary collaboration (Dorst 2015a; Mejía et al. 2023). Such a process would clearly be relevant to research, where unexpected findings, conflicting evidence, or new factors are discovered, initial assumptions and research questions might need to be reconsidered.

From a creativity perspective, problem reframing can be understood as a form of *p-transformational* creativity (Boden 2004). Transformational creativity involves modifying the generative rules that govern a conceptual space, expanding the accessibility of possible solutions for a problem. Problem reframing for an individual precisely does this. Following Boden's distinction between p-creativity (novel to the individual) and h-creativity (novel through history), we use p-transformational to denote transformations when they are novel to the researcher. Similar links between reformulation and creativity have been established across design frameworks. In Gero and Kannengiesser's (2014) Functional-Behaviour-Structure ontology, the most radical reformulation modifies the function space itself. Similarly Grace and Maher (2015) model surprise and reformulation as coupled metacognitive processes. These accounts converge on a shared insight: the creative act in research and design often lies not in finding a better answer, but in finding a better question.

Our work investigates whether Large Language Models (LLMs) can be prompted to produce outputs that researchers find valuable for this kind of reconceptualisation. To directly measure problem reframing with LLMs in a novel research context would require longitudinal observation of researchers over time; instead, we take a first step in examining what makes LLM-generated answers to open questions interesting to researchers who posed them. Through a two-stage user study with 7 research students across diverse disciplines, we first capture each participant's research context through extensive note-taking on previously read papers, then have participants evaluate 27 LLM-generated answers produced by three prompting strategies of varying structure.

By analysing which strategies produce answers that researchers want to pursue further, and by examining how participants rate their top-ranked answers across seven criteria, we identify the characteristics that distinguish answers with potential to support problem reframing from those that do not.

Herein, the main contributions of this paper are as follows:

1. A comparative analysis of three different prompting

strategies - direct question-answering (*direct_strategy*), context-aware question-answering (*all_content*), and a structured approach based on Dorst's Model of problem reframing (*dorst_frame*).

2. An understanding of what makes LLM-generated answers valuable and interesting to researchers.
3. Evidence for two distinct answer profiles, reinforcing and reorienting, and a corresponding set of criteria for evaluating whether LLM-generated answers have reframing potential.
4. An open-source software implementation for an Obsidian Plugin[1] through which we conduct our user study. The plugin allows for the generation and recording of answers across different users, and can be adapted for future investigations of LLM-assisted creative research processes.

Background

This section introduces literature on three aspects of our research. The first subsection introduces Dorst's frame creation methodology, which motivates our *dorst_frame* strategy. The second section discusses how notes serve to capture the context that reframing requires. The third section reviews literature on the capabilities and limitations of LLMs in creative and framing tasks.

Dorst's Frame Creation Methodology

The introduction established that problem reframing is a form of p-transformational creativity. Here, we focus on how reframing has been operationalised in design practice, which directly motivates our prompting strategy.

Schön (1983) established in his work on reflective practices that problem setting is fundamentally important to what information is attended to. Dorst and Cross (2001), for example, provide evidence that design involves co-evolution of the problem and solution spaces with feedback enabling this co-evolution. Building on this, Dorst (2011) identifies frame creation as the core practice of design thinking. In his framework, Dorst (2015b) describes frame creation as a nine-step process that moves beyond generating solutions to creating new approaches to problem situations. This nine-step process is described in (Dorst 2015b, Chapter 4) consisting of the following steps:

1. **Archaeology** - Understand deeply previous attempts to solve the problem and what led to how the problem situation is defined.
2. **Paradox**: Understand what makes the problem hard to solve? Get to the core of the deadlock that prevents the solution of the problem.
3. **Context**: Put away the paradox and shelve the original problem, understand practices of inner stakeholders involved in the problem situation before attempting to solve it.
4. **Field**: Once an overview is achieved, widen the context to form a field, a wide intellectual space.
5. **Themes**: Understand deeper factors of players in the field. Ends with understanding of underlying "universals".
6. **Frames**: Themes, especially those that are shared among many players, might act as a basis for frames, but ideation of a new frame is largely a creative leap.
7. **Futures**: Envision how things might work under a frame and get feedback.
8. **Transformation**: Critically evaluate what frames and solutions are feasible. Unearth changes that are required in the proposed ideas and practices of the participating organisations. Weed out unfeasible frames
9. **Integration**: Integrate frames into broader organisations.

The first three steps deepen understanding of the problem before any frame is proposed. Steps 4 to 5 broaden the perspective, and Step 6 is the creative leap. Steps 7 to 9 concern stakeholders in organisational settings. Since our context and problem are on an individual-level, we adapt the first three steps and step 6 as the basis for our *dorst_frame*.

Notes as Context for Researchers

If deep contextual understanding is essential for problem reframing, how can such an understanding be captured? Research notes (literature notes, field notes, etc.) serve as a primary mechanism through which researchers manage the knowledge they develop. Phillippi and Lauderdale (2018) suggests that field notes, for example, explicitly capture contextual information, including physical setting, social interactions and temporal factors. The Zettelkasten method is another example of a contextualised note system that can sustain creative output, having enabled sociologist Niklas Luhmann to produce over 50 books and 550 articles (Luhmann 1981; Schmidt 2016).

As these examples display, taking notes has high utility for capturing research context, and this has driven the development of digital personal knowledge management tools like Roam Research (Roam Research, Inc. 2026), Notion (Notion Labs, Inc. 2026), and Obsidian (Obsidian Foundation 2026). However, despite the clear importance of notes as context capture mechanisms and the proliferation of digital note-capturing tools, there exists a notable gap in research on how AI systems can effectively utilise researchers' personal notes. While LLM and AI systems have explored the use of context for very specific research problems, for example, in electronic lab notebooks (Fraga et al. 2024; Jalali et al. 2024), and using personal collections of research papers to suggest new research papers (Lee et al. 2024). The same cannot be said for literature notes.

This gap motivates our choice of Obsidian as the platform for our study. Obsidian (Obsidian Foundation 2026) is an open-source, extensible markdown-based software with a robust plugin architecture that allows for custom integration of LLM capabilities while preserving researchers' ownership and control of their notes. Thus, our system exists as a plugin that Obsidian users can use with their personal notes.

[1]The GitHub repository for the paper is: https://github.com/abhinavsood2002/Research-Problem-Reframing-LLMs/

LLMs and Problem Reframing

Our study asks LLMs to generate answers to open research questions that might serve as starting points for problem reframing. This task sits at the intersection of question answering, framing, and creative ideation. LLMs perform well on scientific questions with correct answers (Kon et al. 2025), but our participants' questions require generating new perspectives, not retrieving facts. In computational creativity, framing typically refers to contextualising creative artefacts to shape the audience perception (Charnley, Pease, and Colton 2012; Cook et al. 2019). Our use is closer to Dorst and Schön's sense of generating alternative perspectives on a *problem* to restructure how the researcher thinks about it.

Researchers already use LLMs to seek new information, edit or create new writing, get new ideas, frame research papers and modify or generate data (Liao et al. 2025). When used for ideation, LLMs can directly generate research artefacts like ideas and hypotheses (Wang et al. 2024; Yang et al. 2025; Si, Yang, and Hashimoto 2025). However, such artefacts display consistent limitations: lower technical depth, unrealistic assumptions, and a lack of diversity (Wang et al. 2024; Si, Yang, and Hashimoto 2025). These limitations also apply when LLMs are used for problem reframing. When over 280 designers were performing problem reframing with assistance from LLMs, through 3 different approaches, structured, direct and free-form, the authors found no improvement in the quality of the frames with the LLMs. "The most commonly reported issue with this study's LLM-generated frames was that they were too low-quality to serve as foundations" (Shin et al. 2025). As mentioned earlier, our work does not claim to resolve these known limitations; issues of feasibility and hallucination remain present in our outputs. Instead, we investigate whether structuring LLM prompting around Dorst's reframing methodology affects whether researchers find outputs interesting enough to pursue, and we examine what characteristics distinguish answers with reframing potential from those without.

We focus on measuring interest because, in the context of problem reframing, an answer must first be perceived as worth pursuing by the individual. Otherwise, an uninteresting output will never be engaged with deeply enough to shift how they think about a problem.

System Design

Our system is implemented as an Obsidian plugin that enables researchers to use their existing notes for LLM-generated answers. The system architecture comprises four main components: a React-based frontend integrated into Obsidian, a FastAPI backend server, a vLLM server to run local inference using GPT-oss-120b (medium) (Agarwal et al. 2025) model, and a PostgreSQL database to save user data.

Frontend Architecture

The plugin frontend is built with React and Chakra UI. The interface consists of three primary views:

1. **LoginView** handles user authentication and login to manage data for different participants.
2. **SetupView** enables context configuration by allowing researchers to input their research interest and select relevant notes and PDFs from their Obsidian vault.
3. **FrameBrowserView** displays generated answers. This view is the one that is being displayed when the plugin is embedded in Obsidian's interface in Figure 1. The ranking of answers in Session 2 was done through a ranking modal that appears when the Rank Frames button in the same interface is clicked.

Additionally, we generate markdown so that any tables/formatting can be appropriately rendered. Math equations are rendered through KaTeX.

Backend Architecture

The backend is a FastAPI server. The database schema includes tables for users, notes, PDFs, PDF fulltext, frames, and user contexts.

Frame Generation Strategies The system implements three distinct prompting strategies for frame generation:

1. **Direct Answer Strategy** (*direct_strategy*): Generates answers using only the research interest and research question, without accessing user notes or PDFs. This serves as a baseline.
2. **All Content Strategy** (*all_content*): Provides the LLM with the research interest, research question, and the full content of all selected notes and summarised PDFs as context.
3. **Dorst's Frame Strategy** (*dorst_frame*): Implements a structured four-step sequential process based on Dorst's frame creation methodology (Dorst 2015b). The strategy executes four LLM calls in sequence: **Archaeology**, **Paradox**, **Context**, and **Frames**. Each step builds on previous outputs, with all steps having access to the full content of selected notes and summarised PDFs. Complete prompts for all strategies are provided in the linked GitHub repository.

For each of the strategies that use the notes and PDFs, notes are included verbatim, and PDFs are first conditionally summarised using an LLM prompt that creates 800-1000-word summaries highlighting content relevant to the user's research interest and notes. We use Docling (Team 2024) to parse PDFs into full text. Additionally, P6 and P7 verified that the summaries corresponded to content from the PDFs.

To ensure diverse answers for session two, for each new generation of a question- strategy combination, previously generated answers of that question-strategy combination are also passed to the prompt. The LLM is instructed to pursue a direction that is different from those answers. This, in turn, means that answers are not truly independent of one another; however, strategies and questions still remain isolated.

Study Design

We conducted a two-session user study with 7 research students to investigate how LLMs can support reframing in research contexts. Sessions with participants P1, P2, P3, P6,

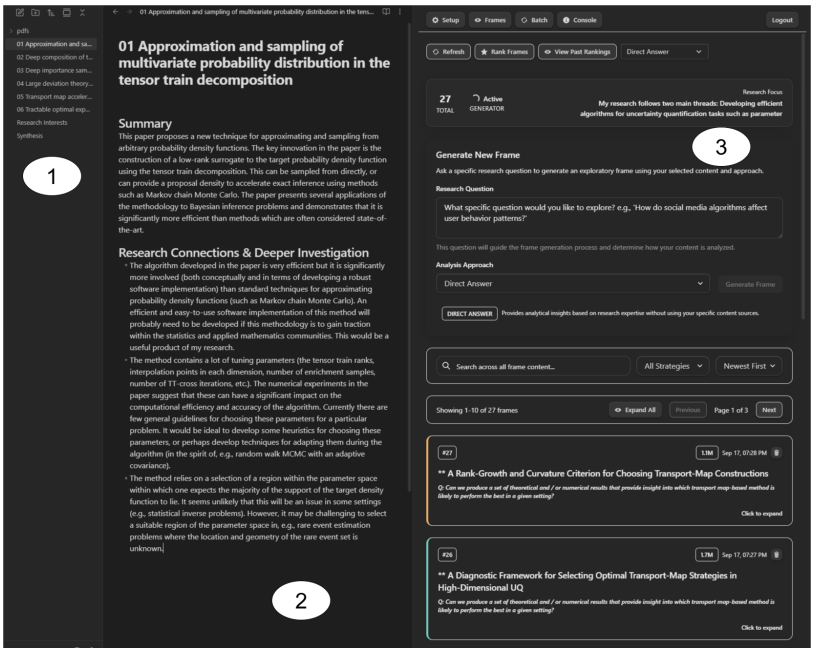

Figure 1: Our Obsidian Plugin. 1: Obsidian file pane representing the different notes the participant has taken, and a *pdfs* folder to store the PDFs of research papers for each note. 2: An example of a participant's note on a research paper. 3: Our plugin that keeps track of different frames generated for each user. Frames are generated in response to a research question and after the context has been set.

and P7 were conducted in person, while sessions with participants P4 and P5 were conducted online. The User Study had human ethics approval from a Human Research Ethics Committee at the University of Sydney. Each participant filled out a consent form agreeing to the collection of audio and computational data.

Participants

We recruited 7 research students through posters, social media and direct outreach to eligible participants. Research students were chosen due to being a convenient demographic. All participants were active student researchers who were engaged with academic literature for their research. Table 1 provides demographic information on our participants. Participants received AU$70 for completing both sessions.

Session 1: Context capture through note-taking

The first session lasted 2 to 2.5 hours and focused on capturing participants' research context through note-taking on previously read papers. The session proceeded in three stages:

1. **Phase 1 Pre-Interview (8-15 minutes):** We conducted a semi-structured interview to understand participants' existing note-taking and use of AI practices. Questions included:
   - Do you use AI (like ChatGPT) in your research? How?
   - Where would you say your research ideas come from?
   - Are there any specialised AI tools that you use for your research work?
   - What kinds of notes do you take while doing research?
   - What is the structure of these notes?
2. **Phase 2 Note-Taking (75-100 minutes):** Participants took notes on research papers they had previously read and had identified as relevant to their research. Participants were given a structured note-taking guide designed to promote more effective note-taking and to explicitly strengthen the connections between ideas, which has been shown to lead to higher-quality notes (Kiewra 1989). The guide describes four types of notes:
   (a) Short summaries highlighting the relevance of the paper to their research
   (b) Connections to the research paper (concepts, questions, or methodologies)
   (c) Aspects that the participant feels warrant deeper investigation
   (d) Surprising or unexpected findings

   Each of these types was elaborated with examples, and the entire guide is available in the linked GitHub repository. While participants were provided the guide, they were instructed to take notes that they thought would be most useful to them and to only follow the guide if they thought it provided them value.
3. **Phase 3: Question Synthesis (15-20 minutes)**: After completing their notes, participants synthesised three research questions based on their note-taking session. These were open questions in their research rather than questions with known answers. Participants rated their satisfaction with the quality of notes produced on a 7-point Likert scale (1 = Very Dissatisfied, 7 = Very Satisfied). This metric was used because a participant's per-

Participant	Gender	Age	Degree / Program (Year)	Research Area / Discipline
P1	Male	18–24	Doctor of Philosophy (Year 1)	Statistics
P2	Female	25–34	Doctor of Philosophy (Year 3)	Human–Computer Interaction
P3	Male	25–34	Doctor of Philosophy (Year 1)	Biomedical Engineering
P4	Male	18–24	MASc, Mechanical Eng. (Year 2)	Process Eng., Commercialisation, Sustainability
P5	Male	25–34	Doctor of Philosophy (Year 3)	Creative AI / Human–AI Co-Creativity
P6	Female	25–34	Doctor of Philosophy (Year 3)	Psychology
P7	Male	35–44	Doctor of Philosophy (Year 3)	Design Innovation, AI

Table 1: Participant Demographics

sonal assessment of their own notes' quality is likely the most important factor while determining the overall quality of the notes (Friedman 2014).

Session 1 - Outcomes

Participants demonstrated diverse note-taking practices prior to taking part in the study. Only P1 maintained an extensive collection of notes on research papers in Obsidian. P3, P4, and P6 mentioned taking limited notes, usually comparing specific biomaterials, commercialisation processes, and impairment modalities, respectively. P4 and P5 did maintain a limited amount of notes in reference management tools like Zotero (Zotero Community 2025) while P3, though documenting extensively in digital and physical diaries, did not systematically take notes on research papers.

After the session, participants were generally satisfied with the quality of their notes. On a 7-point Likert scale, P1, P2, and P3 rated their notes 5/7; P4, P6, and P7 rated their notes 6/7; and P5 rated theirs 7/7. P1 found the guide "representative of the specific kinds of notes I write" and identified the research connections section of the notes to correspond to notes they would personally write the most. P4, P5, P6, and P7 appreciated the structured writing process, with P4 and P5 both noting that engaging in such a process would benefit their thesis writing as they accomplished substantial writing during the session. P2 and P3 preferred to write notes in their own style. P2 primarily on summarising sections relevant to their research with questions interspersed in the notes. P3 recorded key points of the paper relevant to their research interests.

Session 2: Frame Generation and evaluation

The second session occurred after the first session, with a gap of a few minutes to a few days, depending on the participant's schedule. Our system generated 27 answers in response to the three questions provided by each participant. For each question, 3 answers for each strategy were generated.

The produced answers were then further categorised in a 2-step process:

1. First, participants reviewed all 27 generated answers and categorised each as **interesting** or **useless** based on an initial reaction to the answer's potential value to their research and from their interest towards the answer. These answers were provided in an order grouped by question, with the order of answers within each question randomised to prevent bias from strategy. This categorisation produced for each participant 9 answers (capped) that they deemed were interesting, and 18 that were "useless".
2. Participants then ranked their top 9 answers through 36 pairwise comparisons.
3. For the top-3 answers, participants filled out a questionnaire that recorded why that particular answer was interesting to them. This questionnaire contained 8 questions. The first seven were answered on a 7-point Likert scale, and the last one was an open-ended short-form question asking why the answer was interesting. The seven Likert scale questions were:

   - Q1. This answer taught me something about the problem I am working on, which I didn't know before.
   - Q2. This answer helped clear up my thinking about part of the problem I wasn't sure about.
   - Q3. This answer challenges the way I've been thinking about this problem.
   - Q4. This answer connected my problem to something that I hadn't been thinking about before.
   - Q5. This answer helped me feel that I'm on the right track with the way I'm thinking about this problem.
   - Q6. This answer surprised me.
   - Q7. This answer made me think more deeply about the problem.

The session was interactive, participants were encouraged to think their reasoning out loud when possible and were asked what they felt made particular answers interesting/useless. Other questions asked include: Do any of the answers make you rethink your question? For this question, what criteria did you use to categorise answers as interesting?

Results

Participants P1, P3, P4, P5, and P6 reported that the top-3 answers did make them rethink their questions to some extent, indicating the answers provide potential for reframing. To unpack this reframing potential, we first compare how the three prompting strategies performed, then characterise what made answers interesting or useless, and finally examine how participants' ratings across the seven evaluation questions cluster into distinct response profiles.

How Do the Different Prompting Strategies Compare?

Figure 2 presents the overall performance of each strategy in the initial categorisation phase. Answers generated with the *dorst_frame* strategy received the highest rate of interest, with (n=29/189) being classified as interesting. This outperforms *direct_strategy* (n=17/189) generated answers and *all_content* (n=17/189) generated answers.

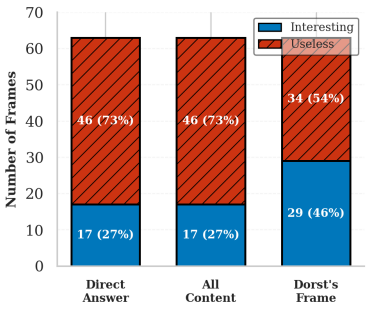

Figure 2: Strategy performance in first-pass categorisation across all participants (N=7)

In the top-3 rankings, Figure 3 shows that *dorst_frame* accounted for more than half of all top-3 ranked frames (n=11, 52%), followed by *direct_strategy* (n=7, 33%) and *all_content* (n=3, 14%).

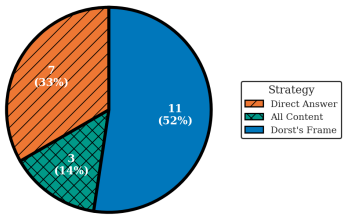

Figure 3: Distribution of prompting strategies among participants' top-3 ranked answers (N=21 total)

While aggregate results favour *dorst_frame*, individual participants exhibited varied preferences as visible in Figure 4. In the initial categorisation phase, six out of seven participants (P1, P2, P4, P5, P6, P7) rated *dorst_frame* answers as interesting more frequently than other strategies. P3 was the exception, showing a preference to *direct_strategy*. The top-3 rankings revealed a more complex pattern. While P1, P4, P5, and P7 maintained their preference for *dorst_frame*, P2 and P3 selected predominantly *direct_strategy* answers.

We argue that this divergence in preference is strongly linked to what participants were looking for in answers. While P1, P4, P5, and P7 mentioned they were looking for specific information related to their questions, P2 and P6 mentioned structure like the answer containing dot points was an important criterion to qualify answers as interesting. P3 described that when they interact with AI, they are looking for a conversation to bounce ideas off of; they actively

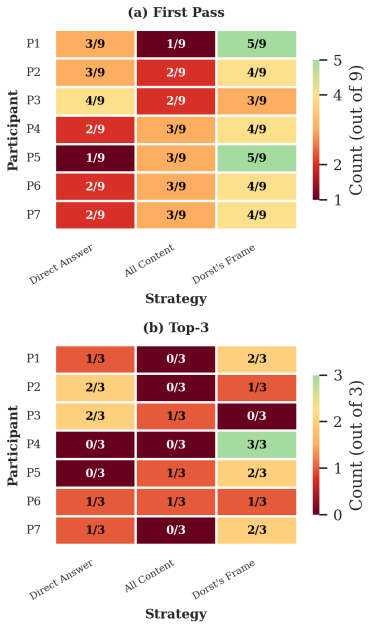

Figure 4: Per-participant strategy performance across evaluation stages. (a) First-pass categorisation showing the distribution of answers' underlying strategies if the answer was rated as interesting (b) Count of each strategy in each participant's top-3 answers

avoid jargon, which was more prevalent in the *dorst_frame* and *all_content* answers. Specifically, for them, interesting frames were described as "they let my creativity flow, if I am sitting there treating it as a person, yeah good point tell me more".

Some participants also observed a qualitative difference in the character of each strategy's output when presented with the answers grouped by strategy at the end of the session, if time permitted. P1 liked the *direct_strategy* answers, with some that were interesting, some less so, and some not making any sense. For *all_content* answers, they were on the whole less interested as "there wasn't enough detail in the answers. The first batch had more nonsense, but more variation". Answers from *dorst_frame* were described as "more similar to the first set, where the number of practically interesting ideas was high. Not as nonsensical as the first set". For P7, *direct_strategy* provided models or frameworks they hadn't heard of before, *all_content* was pulling more direct quotes from the papers, and *dorst_frame* was more assertive and proactive in communicating a particular viewpoint. P6 found it difficult to differentiate the output of the different strategies, though they noted that *direct_strategy* answers were formatted in a way they preferred with more dot points, while simultaneously containing references to research that was "less clear to me".

Taken together, these observations suggest that the three strategies occupied distinct roles: *direct_strategy* acted as a broad but uneven idea generator, *all_content* produced contextually grounded but shallow synthesis, and *dorst_frame* generated more opinionated and structured perspectives that participants could evaluate against their own understanding

of the problem.

What Makes Answers Interesting or Useless?

All participants except P7 ended up marking more than 9 frames as interesting in the first pass and then narrowed them down to their top 9. P7 alternatively expanded their selection from 5 to 9 with a lower standard of interestingness.

Characteristics of Interesting Answers Among the top-3 answers, two groups of interesting answers emerge. One group of answers reinforced participants' existing beliefs and deepened participants' understanding of their question to provide answers that the participants found interesting enough. These answers might not provide as much of an avenue for actual reframing, but still hold value as they might reinforce directions research participants can pursue. These answers typically had a high value for Q2 (cleared my thinking) and Q5 (feel I'm on the right track). This corresponds to what P6 described as a crystallisation effect:

> "I had some of these ideas but it's more crystallised now, sometimes you have these ideas at a very early stage and you don't know how to put it into words but I feel this time it's helping to crystallise what the idea is".

The other set of interesting answers introduced concepts and information that the participant was not thinking or aware of, while being relevant to their own research. Some of these answers proposed complete alternatives. For example, "Replacing petroleum coke with low-sulphur bio-char and by driving the carbothermic reduction in a renewable-electricity-fed plasma furnace" for an answer generated for P4. Others included very specific ideas that the participant had not considered in their context. For example, for P1, "It (the answer) proposed a metric (the concentration index) that I hadn't been thinking about previously but could perhaps be useful". Such answers could be identified by looking for high values of Q1 (answer taught me something I didn't know), Q3 (answer challenges the way I've been thinking) and Q4 (answer connected my problem to something that I hadn't been thinking about) simultaneously. For P5, they perceived answers displayed both these properties.

> "The interesting ones had phrasing that I could understand straightaway at first glance, it combined ideas in a particular way that seemed very familiar to me. It provided solutions that I felt I could work with straightaway and research by myself. It combined concepts in my notes in a surprising way, the terminology was nice, I like how it proposed frameworks, metrics or some computational solution to the problem. It seemed like a new idea hadn't been explored in the space. That was very interesting to me, I would like a copy for myself. "

We call these **reinforcing** and **reorienting** answers, respectively. Participants reported that they had follow-up questions for many of the answers, with that being one of the heuristics for an interesting answer.

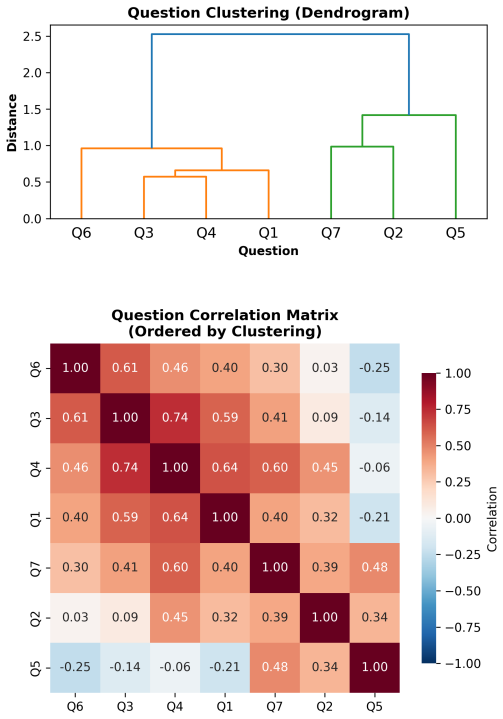

Figure 5: Pairwise Spearman correlations between the seven questions answered for the top-3 frames and agglomerative clustering from these correlations

Characteristics of Useless Answers Useless answers displayed a variety of issues. They misinterpreted terms (P5: "Misinterpreted M-creativity as meta-creativity when I meant model-creativity"); made invalid assumptions (P1: "Some assumed unavailable knowledge to the problem"); did not provide sufficient justification (P4: "Too many calculations, numbers just thrown at face without references"; P2: "Text needs more rationalization, it tries to connect a lot of ideas, but I am not sure how they can practically work that way"); and did not actually answer the question asked (P3: "Did not provide direct answers to the questions and provided additional information I did not want"; P6: "Sometimes I saw responses that do not answer the question"). For P7, many frames misinterpreted the theoretical and philosophical nature of their research papers. P7 noted that "because I had quite a few theoretical/philosophical papers, the AI couldn't really get it. None of them were suggesting anything that was surprising or inspiring". When the LLM did attempt to engage with P7's papers, it often "forced connections between all papers that were not as appropriate", diminishing the credibility of otherwise reasonable outputs.

Which questions correlate with each other?

Figure 5 presents the pairwise correlations for the seven Likert-scale questions across the 21 top-3 answers. These correlations reinforce our 2 distinct kinds of interesting answers. Cluster 1 consists of questions Q1, Q3, and Q4, identifying *reorienting* answers, and cluster 2 consists of Q2 and Q5, identifying *reinforcing* answers. Q6 (surprise)

relates more closely to the questions identifying *reorienting* answers, and Q7 (deepened thinking) to *reinforcing*, but both display additional interesting properties. Surprise was best associated with challenging a participant's thinking (Q3) while making the participant feel they were not on the right track (Q5). We further discuss this in the next section. Q7 and Q2 are the only questions that are positively related to all other questions, with Q7 displaying this property more strongly.

Discussion

Our results provide evidence for two different kinds of answers. **Reinforcing** answers crystallised ideas participants already held (high Q2, Q5), allowing researchers to better prepare their conceptual space of ideas. **Reorienting** answers through introducing new concepts, information, and connections (high Q1, Q3, Q4) pushed participants to consider directions they had not considered, aligning more closely with the conditions for p-transformational creativity. The correlation between Q7 (deepened thinking) and both clusters suggests that both confirmation and challenge can provoke substantive reflection, albeit toward different ends.

As participants had different conceptions of surprise, the role of surprise within this structure requires specific attention. For P3, one of their top three answers included a concept not previously identified in the literature. However, they had recently been discussing this concept with medical students, which left them both surprised and validated when it appeared in the response. P4 and P5 located surprise in unexpected connections: for P4, "something I had no idea about being related to the project", and for P5, the system "bridging two research problems I have been thinking about in a novel and useful way". P6 articulated a more epistemically disruptive form of surprise. For P6, surprise occurs "if I already know something about a topic, and I think I'm correct, but then I read this, and I question if my understanding is correct or not". P7 introduced credibility as a further precondition: surprise means "presenting a really thought-provoking alternative perspective that has credibility".

In related work, Maher, Brady, and Fisher (2013) operationalise surprise as distance between observed and projected attribute values, and Grace et al. (2015) formalise novelty, value, and surprise as independent dimensions of creative design, acknowledging that different kinds of expectation violation can all give rise to surprise. In practice, these models compute surprise as a single score: how much a design deviates from some expected characteristic. Our participant data suggest that in a research context, sources of surprise can be so varied that a single measure might obscure more than it reveals. Instead, more targeted questions that require breaking down surprise by its source, whether from a research connection or from epistemically challenging a participant's understanding, might serve as better models of surprise.

Finally, the gap between *all_content* and *dorst_frame* indicates that merely supplying LLMs with context is not sufficient in producing outputs that researchers can use as bases for reframing. Both the strategies had access to the same notes and PDFs, yet the structure of *dorst_frame* pushed outputs to be described as more interesting. This suggests prompt architectures that mirror problem reframing processes can serve as good candidates for reformulation with AI-generated content.

Limitations

The main limitation of the study is the sample size of participants (n=7). While the quantitative comparisons point out a trend, they lack statistical power. Moreover, problem reframing is an iterative process. Will our participants actually use any of the frames they rated as interesting? Did those frames influence their subsequent research questions? Thus, future work could include providing frames generated in this session to the participants and then using a longitudinal follow-up to see if reframing did occur.

Our system relied on one LLM (GPT-oss-120b). Different LLMs might produce frames with different characteristics with the same strategies. Finally, our study only takes into account part of the modelable research context. That is, it largely focuses on literature notes. As seen in our pre-interviews, researchers keep notes in various forms and expanding what context is processed will be necessary to understand the context required for problem reframing in the various research situations that exist.

Conclusion

This paper investigated whether structuring LLM prompting around Dorst's frame creation methodology produces more interesting answers to researchers' own open questions, and characterised the answer profiles that emerged.

Through a two-stage study with 7 research students, we found that *dorst_frame* answers were rated interesting more frequently and comprised 52% of top-3 ranked answers, suggesting that how an LLM processes context matters more than just the context it receives. We further identified two distinct answer profiles, reinforcing and reorienting. Reorienting answers that correspond to answers with reframing potential correlate with questions Q1 (taught something new), Q3 (challenges thinking), and Q4 (unexpected connection). Thus, these questions serve as a useful proxy for identifying outputs aligned with conditions for p-transformational creativity.

These findings do not resolve known LLM limitations for reframing. Whether the answers participants found interesting would lead to productive reconceptualisation remains an open question.

Author Contributions

Author 1 (AS) conducted the user study, interviewed participants, developed the Obsidian plugin, and prepared the manuscript. Author 2 (SW), Author 3 (CP) and Author 4 (KG) supervised the research and provided direction and feedback for the development of this work.

Acknowledgement

We gratefully acknowledge the Collaborative Intelligence (CINTEL) Future Science Platform for their support towards this work.

References

Agarwal, S.; Ahmad, L.; Ai, J.; Altman, S.; et al. 2025. gpt-oss-120b & gpt-oss-20b model card. arXiv:2508.10925.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Charnley, J. W.; Pease, A.; and Colton, S. 2012. On the notion of framing in computational creativity. In *ICCC*, 77–81.

Cook, M.; Colton, S.; Pease, A.; and Llano, M. T. 2019. Framing in computational creativity-a survey and taxonomy. In *International Conference on Computational Creativity 2019*, 156–163. Association for Computational Creativity (ACC).

Creswell, J. W., and Creswell, J. D. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Dorst, K., and Cross, N. 2001. Creativity in the design process: co-evolution of problem–solution. *Design Studies* 22(5):425–437.

Dorst, K. 2011. The core of 'design thinking' and its application. *Design Studies* 32(6):521–532. Interpreting Design Thinking.

Dorst, K. 2015a. Frame creation and design in the expanded field. *She Ji: The Journal of Design, Economics, and Innovation* 1(1):22–33.

Dorst, K. 2015b. *Frame Innovation: Create New Thinking by Design*. Cambridge, MA, USA: The MIT Press.

Foster, A. 2004. A nonlinear model of information-seeking behavior. *Journal of the American Society for Information Science and Technology* 55(3):228–237.

Fraga, F. P. A.; Poggi, M.; Casanova, M. A.; and Leme, L. A. P. P. 2024. Creating automatic connections for personal knowledge management. *SN Computer Science* 5(5):525.

Friedman, M. C. 2014. Notes on note-taking: Review of research and insights for students and instructors. *Harvard Initiative for Learning and Teaching* 1–34.

Gero, J. S., and Kannengiesser, U. 2014. *The Function-Behaviour-Structure Ontology of Design*. London: Springer London. 263–283.

Grace, K., and Maher, M. L. 2015. Surprise and reformulation as meta-cognitive processes in creative design. In *Proceedings of the third annual conference on advances in cognitive systems ACS*, volume 8.

Grace, K.; Maher, M. L.; Fisher, D.; and Brady, K. 2015. Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation* 3(3-4):125–147.

Jalali, M.; Luo, Y.; Caulfield, L.; Sauter, E.; Nefedov, A.; and Wöll, C. 2024. Large language models in electronic laboratory notebooks: Transforming materials science research workflows. *Materials Today Communications* 40:109801.

Kiewra, K. A. 1989. A review of note-taking: The encoding-storage paradigm and beyond. *Educational Psychology Review* 1(2):147–172.

Kon, P. T. J.; Liu, J.; Ding, Q.; Qiu, Y.; Yang, Z.; Huang, Y.; Srinivasa, J.; Lee, M.; Chowdhury, M.; and Chen, A. 2025. Curie: Toward rigorous and automated scientific experimentation with ai agents. *CoRR*.

Lee, Y.; Kang, H. B.; Latzke, M.; Kim, J.; Bragg, J.; Chang, J. C.; and Siangliulue, P. 2024. Paperweaver: Enriching topical paper alerts by contextualizing recommended papers with user-collected papers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19.

Leedy, P., and Ormrod, J. 2001. *Practical Research: Planning and Design*. Upper Saddle River, NJ and Thousand Oaks, CA: Merrill Prentice Hall and SAGE Publications, 7 edition.

Liao, Z.; Antoniak, M.; Cheong, I.; Cheng, E. Y.-Y.; Lee, A.-H.; Lo, K.; Chang, J. C.; and Zhang, A. X. 2025. LLMs as research tools: A large scale survey of researchers' usage and perceptions. In *Second Conference on Language Modeling*.

Luhmann, N. 1981. Kommunikation mit zettelkästen: Ein erfahrungsbericht. In *Öffentliche Meinung und sozialer Wandel/Public Opinion and Social Change*. Springer. 222–228.

Maher, M. L.; Brady, K.; and Fisher, D. H. 2013. Computational models of surprise in evaluating creative design. In *Proceedings of the fourth international conference on computational creativity*, volume 147, 3. University of Sydney Sydney, Australia.

Mejía, G. M.; Henriksen, D.; Xie, Y.; García-Topete, A.; Malina, R. F.; and Jung, K. 2023. From researching to making futures: a design mindset for transdisciplinary collaboration. *Interdisciplinary Science Reviews* 48(1):77–108.

Notion Labs, Inc. 2026. Notion — all-in-one workspace for notes, tasks, wikis, and databases. https://www.notion.so.

Obsidian Foundation. 2026. Obsidian: A knowledge base that works on local markdown files. https://obsidian.md.

Phillippi, J., and Lauderdale, J. 2018. A guide to field notes for qualitative research: Context and conversation. *Qualitative Health Research* 28(3):381–388. PMID: 29298584.

Roam Research, Inc. 2026. Roam research — a note-taking tool for networked thought. https://roamresearch.com.

Schmidt, J. 2016. Niklas luhmann's card index: Thinking tool, communication partner, publication machine. In *Forgetting Machines. Knowledge Management Evolution in Early Modern Europe*, volume 53.

Schön, D. A. 1983. *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books.

Shin, J.; Polyanskaya, A.; Lucero, A.; and Oulasvirta, A. 2025. No evidence for LLMs being useful in problem reframing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery.

Si, C.; Yang, D.; and Hashimoto, T. 2025. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*, ICLR '25.

Team, D. S. 2024. Docling technical report. Technical report.

Wang, Q.; Downey, D.; Ji, H.; and Hope, T. 2024. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 279–299.

Yang, Z.; Liu, W.; Gao, B.; Xie, T.; Li, Y.; Ouyang, W.; Poria, S.; Cambria, E.; and Zhou, D. 2025. Large language models for rediscovering unseen chemistry scientific hypotheses. In *The Thirteenth International Conference on Learning Representations*, ICLR'25.

Zotero Community. 2025. Zotero.